A METHOD OF IMAGE PRE-ANALYZING OF A MACHINE-READABLE FORM OF NON-FIXED LAYOUT.

Inventors: ZUEV, KONSTANTIN, RU;
GARASHCHUK, RUSLAN, RU.

Current U.S. Class:
Intern'l Class: G04K009/34
Field of Search:

References Cited U.S. Patent Documents

6507671	Jan. 14, 2003	Kagan, et al.	235/383
5864629	Jan. 26, 1999	Wustman	235/383
5822454	Oct. 13, 1998	Rangarajan	399/84

Class G06K009/34

A METHOD OF IMAGE PRE-ANALYZING OF A MACHINE-READABLE FORM OF NON-FIXED LAYOUT.

BACKGROUND OF THE INVENTION

Field of the Invention

The present invention relates generally to an optical character recognition of machine-readable forms, and in particular methods of pre-recognition analysis, especially when fields' and other elements' location is not strongly fixed.

Prior Art.

A widespread method of document pre-recognition processing comprises parsing the image into regions, containing text, and regions containing non-text objects.

Developers of the known systems follow the path of restriction and simplification of document structure, in order to apply existing image structural identification methods.

A US patent #5864629 (January 26, 1999, Wustmann) discloses a method of marking out a particular part of document logical structure for the case, when special characters are present in some document regions and are not present in others. The method is illustrated by the example of regions containing character "$".

The method is not enough universal and can be narrowly applied only for special cases.

A US patent #6507671 (January 14, 2003, Kagan, et al.) discloses a method of marking out filled in data input fields of the document. The method is illustrated by the example of standard machine-readable form, of fixed layout.

The method can be applied only for machine-readable forms of fixed layout. It is rather complicated since the step of logical

structure mark out is combined with the step of intelligent text recognition in data input fields.

Methods using fields fixed layout property of a form are often used in form identification, but have important drawback consisting in fitness thereof for the only specially designed machine-readable form.

For example, the method of US patent #5822454 (October 13, 1998, Rangarajan) discloses the way of pre-processing a document of the only fixed form - "INVOICE". The document has the strongly fixed form either by the fields list, or by their properties and layout.

The shortcomings of the method lay in unfitness for processing any other form of document except the invoice form and also inability to change the fields list or overcome spatial deflection of fields.

SUMMARY OF THE INVENTION

One or more graphic and/or text elements in the image of machine-readable form are preliminarily assigned to be used further as reference points for searching data input fields. Information about reference points spatial characteristics and data input fields relative location is placed in special storage means.

A machine-readable form image of a filled paper form is parsed into regions containing data input fields, connected regions, lines and other kinds of objects. Then the reference points' location is defined on the form image. In a case of multiple fields identification result the most closely matching field is selected. The fitness is measured in relation with spatial characteristics mainly.

BRIEF DESCRIPTION OF THE DRAWING

Fig 1 shows a document with three assigned reference points.

Fig 2 shows a spatial binding of a data input field with one of the reference points.

DETAILED DESCRIPTION OF THE INVENTION

A data conversion from paper media into electronic form is processed now by optical character recognition (OCR) systems. Nowadays OCR systems can reliably enough recognize documents of high and middle quality level, typed by any of the standardized font. This provides to convert a large amount of text data, with sometimes no manual control. But a character-by-character conversion is deficient for data transformation into electronic form. That's why the main common target of the prior systems is the document logical structure identification, i.e. evident identification of document's elements properties - title, data input fields, etc.

The main difficulty herein is due to deficient of the existing paper documents regulation. Even if the list of document fields if strongly prescribed (defined, fixed), as, for example, in money order, invoice, tax return or other account document form, the fields location areas are specified only approximately with some tolerance.

Additionally the exact fields location can be hardly assured in a mass typing. Thus, the problem of printed documents logical structure identification comes to the document pre-recognition process.

Spatially the document is formed of structural elements as text regions, various graphic objects and separating lines. Some text regions relate to document fields. Other text regions, graphical objects and separating lines relate to form.

The spatial structure properties can be described by means of spatial and parametric characteristics of structural elements:

- structural element absolute layout restrictions,
- structural element relative layout restrictions,
- structural element dimensions variety restrictions,
- structural element is optional.

An absolute structural element location is typical for questionnaires and other standard forms. At the same time the relative location thereof is fixed as well. But the problem of accurate whole-page location as distortion, shift and turn caused by scanning remains unsolved. In other cases restrictions of absolute location can be used for elimination fully incorrect versions of structural elements identification.

The relative limitations may be divided into two groups:

- qualitative restrictions,
- quantitative restrictions.

The qualitative restrictions show the general type of relationship, for example, that one element is located upper or lower then the other one. The qualitative restrictions specify coordinates of the region, where an element is located relatively to another element. Usually, the qualitative restrictions indicate small mutual location deflections as well. In other cases the qualitative relationship is used.

Sometimes a form structural requirements are not strong enough and do not fix the mutual location of elements, thus permitting the structural variations.

A most widespread type of structural variations is an absence of one or more structural element in the image. Usually this happened due to the optional nature of its presence.

Besides this some elements may also degrade at scanning, that makes their identification impossible. The examples of elements

of this kind are separating lines or text elements of the form, typed in small font. Nevertheless the use of such elements is important since they can additionally specify the other elements' location.

The technical result of the present invention consists in improving an ability to process pre-recognition of machine-readable forms of non-fixed layout.

The mentioned shortcomings greatly reduce the use of known methods to find and mark out data input fields in the image of the form of non-fixed layout.

All known methods are unfit to achieve the declared technical result.

The declared technical result is achieved as follows.

One or more graphic and/or text elements (1) are preliminarily assigned on a form to be used further as reference points for searching data input fields (2).

The assigned elements (1) should be reliably identified on the scanned image.

Information about reference points spatial characteristics and data input fields (2) relative location is placed in a storage means, one of the possible embodiments of which is the form model description.

Additionally, the form model description contains parametrical data about data input fields, for example, the length of the field, the range of permissible values etc.

The graphic image of the machine readable form, after possible skew, distortion and noise elimination, is parsed into regions containing data input fields, connected regions, lines, other objects.

Then the location of objects previously assigned as reference points is determined. In the case the reference point is a text object its contents is additionally recognized.

The data input fields location is determined relatively to one or more reference point. If a full or a part covering thereof happens to similar fields and/or fields of the same type, the most suitable one of them should be chosen. The suitability is measured by the closeness of spatial characteristics thereof to the model. The additional parametrical information about data input fields may also be added if necessary.

The identification is processed by means of setting up and accepting hypothesis about the field.

A reliability estimation of the identification versions, i.e. hypotheses, is processed as follows.

To compare and combine estimations of various structural elements, it is necessary to reduce them to a common scale. The reliability estimation is usually interpret as conditional probability likelihood estimation:

$Q_t \sim p(t|I)$, where t - is a structural element, I - is an image.

On the basis of probabilistic interpretation of reliability estimation the following can be accepted:

$p(N|I) = p(N|t_1,\ldots,t_n)p(t_1|I)\ldots p(t_n|I)$,

where $p(t_i|I)$ - is a probability of I-th element that is a component of composite element N,

$p(N|t_1,\ldots,t_n)$ - conditional probability likelihood estimation that the list of sub-elements comprises composite element N.

The identification reliability estimation of composite element is calculated according to the following rule:

$Q = Q_R Q_1 \ldots Q_n$,

Where $Q_i \sim p(t_i|I)$ - is reliability estimation of i-th element or the composite element N.

$Q_R \sim p(N|t_1,\ldots,t_n)$ - a relations estimation.

The value of $p(N|t_1,\ldots,t_n)$ corresponds with relations Q_R, which restrict sub-elements in the composite element. The relations *should model a conditional reliability density function* $p(N|t_1,\ldots,t_n)$. The said problem is solved using non-clear logic toolkit in the probabilistic interpretation. Logic operations are performed according to the following rules:

operation '$\wedge$' is performed as : $A \wedge B \rightarrow a * b$;

operation '$\vee$' is performed as : $A \vee B \rightarrow a + b - a * b$;;

operation '$\neg$' is performed as : $\neg A \rightarrow 1 - a$.

The conditional reliability density function is formed via partly linear approximation.